Cocoa Price Prediction Based on Time Series Models*

Yizhuo Liu, Leo Kaixuan Cheng, Haobo Ren, Betty (Ruoran) Li

April-04-2024

Contents

*Code and data are available at: https://github.com/YizhuoLiu/ts_forecast_cocoa_price

1 Introduction

Cocoa is an important agricultural product, especially for some West African countries like Ghana and Côte d'Ivoire (Siaw et al. (2021)). These two nations produce more than half of the world's cocoa supply. Therefore, changes in cocoa prices have very important effects on their economies and the income of farmers. This study aims to find the key factors that influence cocoa futures prices. The primary purpose is to develop forecasting models that help predict future price movements.

Cocoa prices often change due to external shocks. Climate change is one major factor. Studies show that abnormal weather can harm cocoa trees and reduce production. This supply shortage often causes prices to rise sharply. For example, in Ghana and Côte d'Ivoire, rising temperatures and irregular rainfall have already created serious problems for cocoa farmers (Siaw et al. (2021)). Also, inflation and changes in exchange rates can change production costs and influence market demand. When inflation increases in cocoa-importing countries or the U.S. dollar strengthens, cocoa prices become more unstable (Addai et al. (2020)). Another important factor is the disease of the cocoa trees, which are susceptible to pests and viruses, such as the swollen shoot virus. These diseases reduce the number of healthy trees, leading prices to increase due to lower production (Tabe-Ojong et al. (2024)). All these factors are connected and must be considered together when modeling cocoa prices.

In this report, we will use time series forecasting models, including ARIMA and linear regression with exogenous variables. These models will help identify patterns of cocoa price predictions. The analysis is based on historical data and includes climate, economic, and agricultural indicators.

There are some key challenges in this study. The data may be missing or incomplete. Also, some time series are non-stationary, which means their trends change over time. Finally, combining different types of data can be very complex. To solve these problems, we will use careful preprocessing steps and test different models to choose the most accurate and reliable one.

2 Literature Review

Time series forecasting has been widely used for modeling commodity prices. Classical models like ARIMA and GARCH are popular due to their simplicity. However, these models assume the linearity and stationarity of the series, which do not hold for all real-world data. An increasing number of studies have started exploring machine learning methods, which can better capture non-linear patterns of the data.

A study in 2010 used univariate ARIMA models to predict cocoa bean prices (Kamu et al. (2010)). The study tested ARIMA on monthly cocoa price data. It shows that ARIMA is very effective in capturing seasonality and trends. However, it cannot directly include the influence of the explanatory variables. Its performance is not very good when data is highly influenced by other factors. Our study uses ARIMA as our preliminary model, and given the problem mentioned above, we also built an ensemble model by combining the predictions from ARIMA and linear regression to improve the forecast. ARIMA captures trends and patterns over time, while linear regression focuses on relationships of cocoa prices with external variables. By averaging the predictions of these two models, the ensemble can reduce errors and balance the strengths of both approaches.

As mentioned before, this study also uses a linear regression model on its own to forecast cocoa futures prices. Linear regression is simple and allows the participation of external variables. Although it may not capture complex time series, it intuitively shows how each factor influences the price directly.

Another model used in this study is ARIMAX, which extends the ARIMA model by including external variables. A recent study used ARIMAX to forecast California's energy consumption using population, production levels, energy prices, and so on (Moslemi et al. (2024)). Their results show that ARIMAX gives more accurate forecasts than ARIMA, especially when external variables highly influence the response variable. This supports the idea that adding predictors can improve model performance. In our study, ARIMAX is used to include temperature, precipitation, inflation rate and disease reports in cocoa-producing countries. This is better for modelling and forecasting real-world data than the univariate ARIMA approach.

In this study, we explore a convolutional neural network (CNN) approach to forecast cocoa futures prices. CNNs are well-known in image and speech recognition, and they are increasingly used for time series forecasting. Another study proposed a WaveNet-inspired CNN model for conditional time series forecasting. Their model uses dilated convolutions to learn long-term dependencies and can make forecasts based on multiple related time series. CNN uses fewer parameters than RNNs and is faster to train (Borovykh et al. (2018)). We applied a simplified version of the temporal CNN to test whether it can capture the cocoa prices when it's affected by temperature.

3 Data

4 Methodology

We split our data into a training set and a test set using a 80% - 20% ratio. Besides the daily price data, we extract a monthly and yearly time series data by averaging all price in one month or one year to see if our models are robust enough to train on those data. For NA and missing values in the dataset, we interpolate the data directly.

4.1 Preliminary Model - ARIMA

First of all, we have chosen ARIMA to be our Preliminary model. As mentioned in the Literature Review, ARIMA seems to be one of the best models when it comes to predicting the price of Cocoa and our initial plot of cocoa price doesn't seem to be stationary. In order for our ARIMA model to perform better, we need to find the best degree of difference and dependence order for our data. Since our data is non-stationary we apply a first difference ($d = 1$) to our data and re-assess. If the data is stationary, we continue to find the AR and MA Orders (p and q), and if the first difference isn't enough to make the data stationary we will introduce a second difference into our model and so on. After determining the degree of difference, we plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). A sharp cut-off in the PACF suggests the appropriate AR order (p) for our differenced data. And a sharp cut-off in the ACF is used to determine the MA order (q). Using the method above, we can find ARIMA models with appropriate degree of difference and dependence order. In general, we will write the model as

$$\Phi(B)(1-B)^d y_t = \Theta(B)\varepsilon_t$$

where ϵ_t is the white noise sequence. In practice, we will use the function (ARIMA) in R to generate an ARIMA model for forecasting.

4.2 Model Ensemble - ARIMA + Linear Model

As good as our time domain approach time series model might be, it lacks consideration of influencing factors in Cocoa price such as climate change, disease and exchange rate. So to compensate for the lack of factors we decided to introduce an ensemble model which not only contains an ARIMA model, but also a linear model which includes factors influencing cocoa price as predictors. We will do the same as described in our preliminary model for the ARIMA part in the ensemble model. The linear regression model is going to be a little more complicated. Naturally we choose our response variable to be the price of the Cocoa, and our predictors to be factors that influence Cocoa price the most. This will give us a general equation for our linear model.

$$y_t = \sum_{i=1}^{N} \beta_i x_t^i + \epsilon$$

where x_i^t are predictors i at time t and ϵ is the noise. The parameters of the linear model will be fitted using least square estimation (LSE). Then we will drop predictors using p-value and T-test (with alpha value being

0.05) to reduce redundancy in our model. Last but not the least, we need to combine the 2 models, because It is hard to evaluate fairly which model should have a higher weight. We decided to use a simple average approach to the ensemble model. This give us the equation for our model:

$$y_t = \frac{y_t^{\text{ARIMA}} + y_t^{\text{LM}}}{2}$$

The ensemble model is useful since our linear model and ARIMA model capture different aspects of the data and we want to combine their strengths.

4.3 ARIMAX

The dependence of the price on other factors such as weather and diseases may be important. Therefore, we choose to use a conditional ARIMA model which is called autoregressive integrated mean average with exogenous variables (ARIMAX). It takes care of the dependence of the time series price with other time series with an expression:

$$y_t = \Phi(B)(1-B)^d y_t = \Theta(B)\epsilon_t + \beta^T \boldsymbol{x}_t$$

where ϵ_t is a white noise sequence and x_t is our exogenous variable. If we expand this, we can get

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \epsilon_t$$

This is a kind of conditional ARIMA, where our time series of interest is conditioned on other variables that are potentially influential. With that, we may be able to predict the price more accurately using the other variables. In practice, we use price as the y variable and we treat everything else in the dataset as the X variable. The ARIMA parameters are chosen based on inspection and model selection techniques such as AIC and MSE. The model is then fitted using a standard ARIMA technique.

4.4 Linear Regression on Lags

During our investigation on background information about Cocoa price, the most important factor which influenced the Cocoa price was the Cocoa price before. So we decided to introduce a linear regression model which contains the price of Cocoa leading to the prediction day. The method for building the linear regression on lags model is no different than our linear regression model in ensemble model. But this time we not only have final predictors in our ensemble model, but also price of Cocoa 1 day before to 10 days before. Then we will use the p-value and t-test again to determine which predictors to keep. The formula for this linear regression model is given by

$$y_t = \sum_{t=1}^{10} \alpha_t y_t + \sum_{i=1}^{N} \beta_i x_t^i + \epsilon$$

4.5 Temporal CNN

Convolutional neural network (CNN) is a popular machine learning model when dealing with image and sequential data. More precisely, when it comes to univariate time series data in our case, we will perform a 1d convolution, which basically convolves some kernels with the predefined receptive field before time step t. After that, take the loss function as that distance between the predicted time series and our training data then backpropagate the gradient as in usual machine learning practice. CNN can easily extract temporal dependence inherently. We are using dilation on CNN to better extract the feature.

More specifically, we are using an encoder-decoder structure. We encode the time series and the corresponding exogenous time series using temporal CNN into some hidden state. We treat different features of the time series as different channels. In this process, we also utilize recurrent neural networks (RNN) to make the network respect the long term dependency. The encoder consists of many residual blocks. Then, using multi-layer perceptrons (MLP), we construct a decoder to predict future values conditioned on the past time series as well as the exogenous variables. The model is fully differentiable so we optimize it using an AdamW optimizer with appropriate hyperparameters tuned by the validation error on a test set splitted from the dataset.

5 Results and Forecasting

5.1 Preliminary Model

5.2 Model Ensemble - ARIMA + Linear Model

5.3 ARIMAX

The ARIMAX model is fitted using daily data as well as the monthly data with parameter chosen automatically. The inference results on the monthly and daily data can be seen from the diagram below. We can see that the result is relatively bad although we take into consideration other factors. It is similar to the results with a vanilla ARIMA model: the predicted time series becomes straight as a line, the error bar is enormous compared to the scale of the time series. The overall upward trend is correct except for the drastically increasing. This shows that the model is very uncertain on its prediction, and simply cannot fit the data because of its complexity. This suggests that ARIMAX is not a good choice for this data.

The final ARIMAX model chosen for daily data based on AIC is ARIMA(0,1,0), showing that it fits very bad on the data since the data is highly non-stationary and the bset ARIMA model even does not depend on the past. For monthly data, the model is ARIMA(0,1,1).

5.4 Linear Regression on Lags

5.5 Temporal CNN

It takes the CNN ~1000 epochs to converge to a low loss. The fitted time series on the training and test set are shown. We can see that the model performs extremely well on the training set, showing that the expressiveness of the model is sufficient to capture the training data. We can see that the trend of the time series is correctly predicted by the CNN, but there is still some deviations from the true test set. The validation MSE loss on the test set is also large. A potential reason is that the model is too complex and it overfits the given training set even though we apply a standard L^2 regularization on its weights. Overly complex model may not be suitable for a time series prediction without good regularization. The loss curve is also shown below.

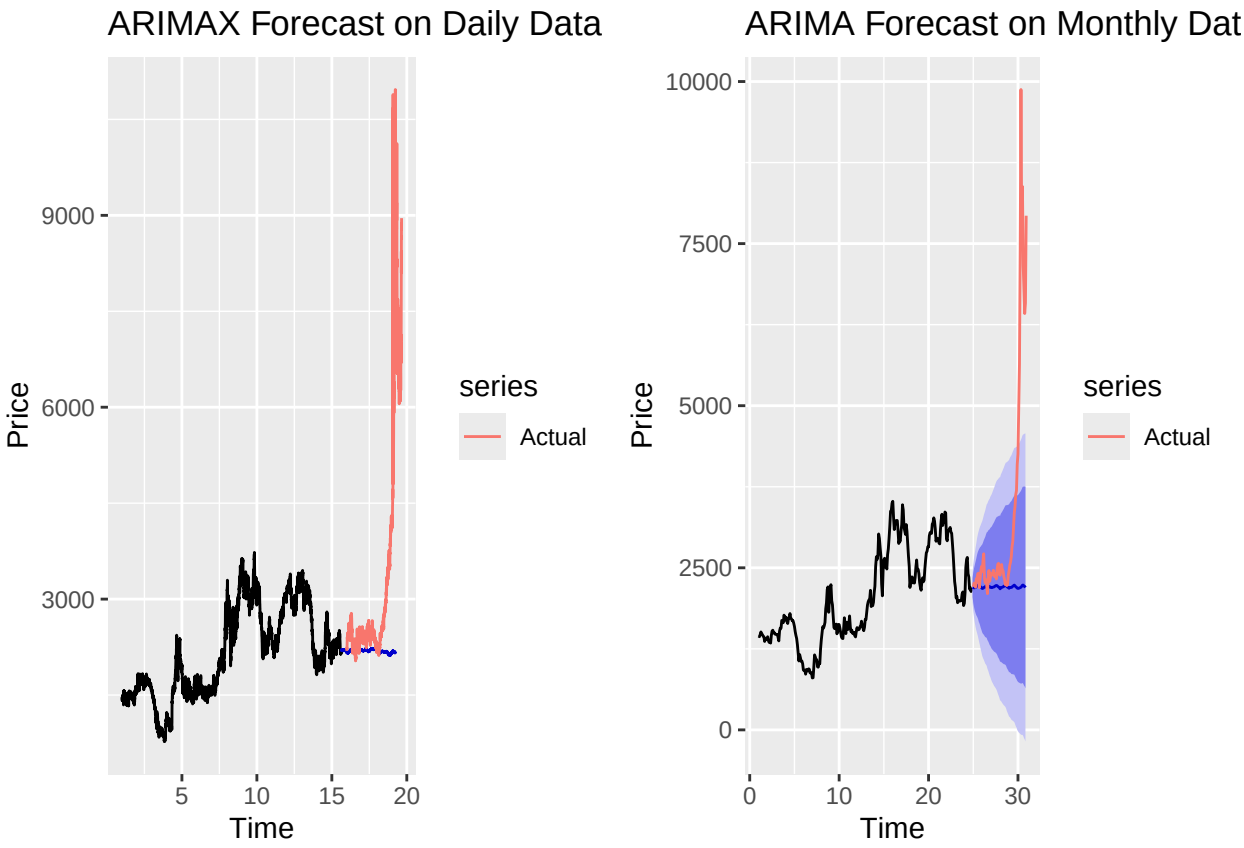

Figure 1: ARIMAX model fitted on daily data.

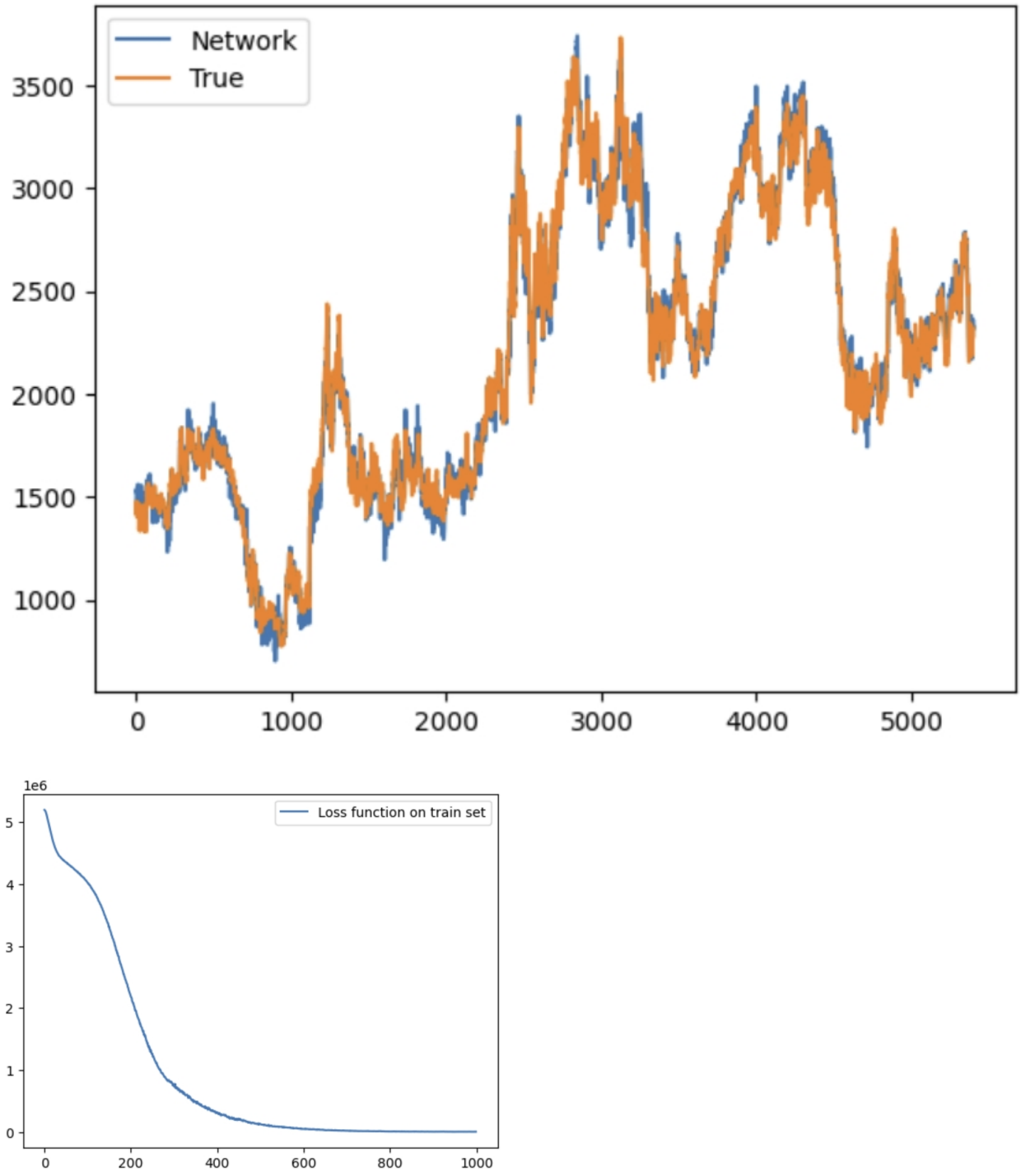

Figure 2: Side-by-side comparison of image A and image B.

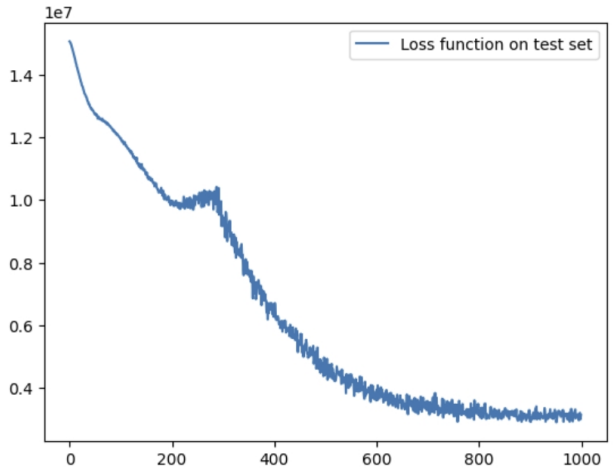

Figure 3: Side-by-side comparison of image A and image B.

6 Discussion

Overall the traditional time series prediction methods all perform relatively bad on the cocoa global price dataset. This shows that this data has an inherently complex nature, and many models are bad at capturing the pattern of the data. Another factor that may be problematic is the stationarity of the time series, which is hard to achieve. Also, traditional models such as ARIMA may fail because they don't take into consideration the exogenous variables such as weather and inflation. To resolve this, we found that the linear model trained on lags achieved a satisfying performance on both the training set and the generalization to the test set. Theoretically, temporal CNN will also perform well on this the performance of the linear model. However, due to issues like overfitting, the CNN model is not good as the linear model when predicting the future.

Although the vanilla ensemble model does not perform well, we may use the idea to combine two equally good models and tune the weight between the outputs of the two models to gain a better result. Or we can use another function to dynamically combine the results given by two or more models.

Data quality may also be important. We may incorporate more relevant data to train our models.

Recently there are works on utilizing diffusion models for time series prediction (Yuan & Qiao, 2024), which achieve state-of-the-art performance. Using that, we may be able to fit a robust and well-generalized model on this complex dataset conditioned on other factors. The flexibility of the proposed model enables us to do that.

7 Conclusion

Reference

Addai, B., Gyimah, A. G., & and, K. P.-A. (2020). Exchange rate regimes and global cocoa trade: To float or to peg? *Cogent Economics & Finance*, *8*(1), 1719593. https://doi.org/10.1080/23322039.2020.1719593

Borovykh, A., Bohte, S., & Oosterlee, C. W. (2018). *Conditional time series forecasting with convolutional neural networks*. https://arxiv.org/abs/1703.04691

Kamu, A., Ahmed, A., & Yusoff, R. (2010). Forecasting cocoa bean prices using univariate time series models. *Journal of Arts Science & Commerce*, *1*, 71.

Moslemi, Z., Clark, L., Kernal, S., Rehome, S., Sprengel, S., Tamizifar, A., Tuli, S., Chokshi, V., Nomeli, M., Liang, E., Bidgoli, M., Lu, J., Dasaur, M., & Hodgett, M. (2024). *Comprehensive forecasting of california's energy consumption: A multi-source and sectoral analysis using ARIMA and ARIMAX models*. https://arxiv.org/abs/2402.04432

Siaw, D., Mushi, V., & Tindana, P. (2021). *Effect of climate change on the price of cocoa a case study of ghana and cote d'ivoire and the intercontinental exchange (ICE)* [PhD thesis]. https://doi.org/10.13140/RG.2.2.30741.40160

Tabe-Ojong, M. P. Jr. et al. (2024, May 8). *Soaring cocoa prices: Diverse impacts and implications for key west african producers*. International Food Policy Research Institute (IFPRI). https://www.ifpri.org/blog/soaring-cocoa-prices-diverse-impacts-and-implications-key-west-african-producers/

Yuan, X., & Qiao, Y. (2024). *Diffusion-TS: Interpretable diffusion for general time series generation*. https://arxiv.org/abs/2403.01742